Attorney Docket No.: 018547-034810US
Client Reference No.: 3079A

PATENT APPLICATION

Sub B1

COMPUTER-AIDED VISUALIZATION OF EXPRESSION COMPARISON

Inventor(s): David H. Mack, a citizen of The United States, residing at
2076 Monterey Avenue
Menlo Park, CA 94025

Kurt Gish, a citizen of The United States, residing at
953 Helen Avenue, Apt. 7
Sunnyvale, CA 94086

David Balaban, a citizen of The United States, residing at
7127 Glenview Drive
San Jose, CA 95120

Elina Khurgin, a citizen of The United States, residing at
22999 Voss Avenue
Cupertino, CA 95014

Josie Dai, a citizen of China, residing at
4009 Higuera Road
San Jose, CA 95148

Jim Snyder, a citizen of The United States, residing at
321 Fulton Street
Palo Alto, CA 94301

Assignee: Affymetrix, Inc.
3380 Central Expressway
Santa Clara, CA 95051

Entity: Large

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 650-326-2400

PATENT

Attorney Docket No. 018547-034800US / Client 3079

'1-05
B2

BACKGROUND OF THE INVENTION

The present invention relates to the field of computer systems. More specifically, the present invention relates to computer systems for visualizing analysis results.

Devices and computer systems for forming and using arrays of materials on a substrate are known. For example, PCT Publication No. WO 92/10588, incorporated herein by reference for all purposes, describes techniques for sequencing or sequence checking nucleic acids and other materials. Arrays for performing these operations may be formed according to the methods of, for example, the pioneering techniques disclosed in U.S. Patent No. 5,143,854 and U.S. Patent No. 5,593,839 both incorporated herein by reference for all purposes.

According to one aspect of the techniques described therein, an array of nucleic acid probes is fabricated at known locations on a substrate or chip. A fluorescently labeled nucleic acid is then brought into contact with the chip and a scanner generates an image file (which is processed into a cell file) indicating the locations where the labeled nucleic acids bound to the chip. Based upon the cell file and identities of the probes at specific locations, it becomes possible to extract information such as the monomer sequence of DNA or RNA. Such systems have been used to form, for example, arrays of DNA that may be used to study and detect mutations relevant to cystic fibrosis, the P53 gene (relevant to certain cancers), HIV, and other genetic characteristics.

Computer-aided techniques for monitoring gene expression using such arrays of probes have also been developed as disclosed in U.S. Patent Application No. 08/828,952 (Attorney Docket No. 16528X-028900US) and PCT Publication No. WO 97/10365 (Attorney Docket No. 16528X-017110PC), the contents of which are herein incorporated by reference. Many disease states are characterized by differences in the expression levels of various genes either through changes in the copy number of the genetic DNA or through changes in levels of transcription (*e.g.*, through control of initiation, provision of RNA precursors, RNA processing, etc.) of particular genes. For example, losses and gains of genetic material play an important role in malignant

transformation and progression. Furthermore, changes in the expression (transcription) levels of particular genes (*e.g.*, oncogenes or tumor suppressors), serve as signposts for the presence and progression of various cancers.

It is desirable to identify genes having expression levels relevant to diagnosis of a diseased state by analyzing the expression levels of large numbers of genes in both diseased and normal individuals. Methods for collecting the expression level information have been developed. However, the user interfaces for gene expression monitoring systems that have been developed until now are designed to clearly present the expression of particular pre-selected genes. A user seeking to identify, e.g., an oncogene or a tumor suppressor gene, must individually review the expression level of large numbers of genes and compare the expression levels between diseased and normal individuals. What is needed is a user interface that takes advantage of collected gene expression information to help the user to identify particular genes of interest.

SUMMARY OF THE INVENTION

The present invention provides innovative systems and methods for visualizing information collected from analyzing samples. The samples may include nucleic acids, proteins, or other polymers. Gene expression level as determined from analysis of a nucleic acid sample is one possible analysis result that may be visualized. In one embodiment, a computer system may display the expression levels of multiple genes simultaneously in a way that facilitates user identification of genes whose expression is significant to a characteristic such as disease or resistance to disease. Additionally, the computer system may facilitate display of further information about relevant genes once they are identified.

A first aspect of the invention provides a computer-implemented method for presenting expression level information as collected from first and second samples. The method includes steps of: displaying a first axis corresponding to expression level in the first sample, and displaying a second axis substantially perpendicular to the first axis, the second axis corresponding to expression level in the second sample. The method further includes a step of: for a selected expressed sequence, displaying a mark at a position. The position is selected relative to the first axis in accordance with an expression level of the selected expressed sequence in the first sample and relative to the second axis in accordance with an expression level of the selected expressed sequence in the second

sample. A particularly useful application is displaying many marks simultaneously for many selected genes to discover which ones of the selected genes may be relevant to the characteristic.

A second aspect of the invention provides a computer-implemented method of presenting sample analysis information. The method includes steps of: displaying a first axis corresponding to a concentration of a compound in a first sample as determined by monitoring binding of the compound to a selected polymer having binding affinity to the compound, and displaying a second axis substantially perpendicular to the first axis. The second axis corresponds to a concentration of the compound in the second sample as determined by monitoring binding of the compound to the selected polymer. The method further preferably includes a step of displaying a mark at a position. The position is selected relative to the first axis in accordance with the concentration in the first sample and relative to the second axis in accordance with the concentration in the second sample.

A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the attached drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates an example of a computer system that may be used to execute software embodiments of the present invention.

Fig. 2 shows a system block diagram of a typical computer system.

Fig. 3 illustrates an overall system for forming and analyzing arrays of polymers including biological materials such as DNA or RNA.

Fig. 4 is an illustration of an embodiment of software for the overall system.

Fig. 5 shows a flowchart of a process of monitoring the expression of a gene by comparing hybridization intensities of pairs of perfect match and mismatch probes.

Fig. 6 shows a screen display illustrating gene expression levels for multiple genes as collected from both normal and diseased tissue.

Figs. 7A-7B show screen displays illustrating information about a particular gene selected from the display of Fig. 6.

DESCRIPTION OF SPECIFIC EMBODIMENTS

The present invention provides innovative methods of monitoring visualizing gene expression. In the description that follows, the invention will be described in reference to preferred embodiments. However, the description is provided for purposes of illustration and not for limiting the spirit and scope of the invention.

Fig. 1 illustrates an example of a computer system that may be used to execute software embodiments of the present invention. Fig. 1 shows a computer system 1 which includes a monitor 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have one or more buttons such as mouse buttons 13. Cabinet 7 houses a CD-ROM drive 15 and a hard drive (not shown) that may be utilized to store and retrieve software programs including computer code incorporating the present invention. Although a CD-ROM 17 is shown as the computer readable medium, other computer readable media including floppy disks, DRAM, hard drives, flash memory, tape, and the like may be utilized. Cabinet 7 also houses familiar computer components (not shown) such as a processor, memory, and the like.

Fig. 2 shows a system block diagram of computer system 1 used to execute software embodiments of the present invention. As in Fig. 1, computer system 1 includes monitor 3 and keyboard 9. Computer system 1 further includes subsystems such as a central processor 50, system memory 52, I/O controller 54, display adapter 56, removable disk 58, fixed disk 60, network interface 62, and speaker 64. Removable disk 58 is representative of removable computer readable media like floppies, tape, CD-ROM, removable hard drive, flash memory, and the like. Fixed disk 60 is representative of an internal hard drive or the like. Other computer systems suitable for use with the present invention may include additional or fewer subsystems. For example, another computer system could include more than one processor 50 (*i.e.*, a multi-processor system) or memory cache.

Arrows such as 66 represent the system bus architecture of computer system 1. However, these arrows are illustrative of any interconnection scheme serving to link the subsystems. For example, display adapter 56 may be connected to central processor 50 through a local bus or the system may include a memory cache. Computer system 1 shown in Fig. 2 is but an example of a computer system suitable for use with the present invention. Other configurations of subsystems suitable for use with the present invention will be readily apparent to one of ordinary skill in the art. In one

embodiment, the computer system is an IBM compatible personal computer.

The VLSIPS™ and GeneChip™ technologies provide methods of making and using very large arrays of polymers, such as nucleic acids, on very small chips. *See* U.S. Patent No. 5,143,854 and PCT Patent Publication Nos. WO 90/15070 and 92/10092, each of which is hereby incorporated by reference for all purposes. Nucleic acid probes on the chip are used to detect complementary nucleic acid sequences in a sample nucleic acid of interest (the "target" nucleic acid).

It should be understood that the probes need not be nucleic acid probes but may also be other receptors, such as antibodies, or polymers such as peptides. Peptide probes may be used to detect the concentration of other peptides, proteins, or other compounds in a sample. The probes must be carefully selected to have bonding affinity to the compound whose concentration they are to be used to measure.

In one embodiment, the present invention provides methods of visualizing information relating to the concentration of compounds in a sample as measured by monitoring affinity of the compounds to probes. In a particular application, the concentration information is generated by analysis of hybridization intensity files for a chip containing hybridized nucleic acid probes. The hybridization of a nucleic acid sample to certain probes may represent the expression level of one more genes or expressed sequence tags (ESTs). The expression level of a gene or EST is herein understood to be the concentration within a sample of mRNA or protein that would result from the transcription of the gene or EST.

Expression level information visualized by virtue of the present invention need not be obtained from probes but may originate from any source. If the expression information is collected from a probe array, the probe array need not meet any particular criteria for size and density. Furthermore, the present invention is not limited to visualizing fluorescent measurements of bondings such as hybridizations but may be readily utilized to visualize other measurements.

Concentration of compounds other than nucleic acids may be visualized according to one embodiment of the present invention. For example, a probe array may include peptide probes which may be exposed to protein samples, polypeptide samples, or other compounds which may or may not bond to the peptide probes. By appropriate selection of the peptide probes, one may detect the presence or absence of particular compounds which would bond to the peptide probes.

For purposes of illustration, the present invention is described as being part of a system that designs a chip mask, synthesizes the probes on the chip, labels nucleic acids from a target sample, and scans the hybridized probes. Such a system is set forth in U.S. Patent No. 5,571,639 which is hereby incorporated by reference for all purposes. However, the present invention may be used separately from the overall system for analyzing data generated by such systems, such as at remote locations, or for visualizing the results of other systems for generating expression information, or for visualizing concentrations of polymers other than nucleic acids.

Fig. 3 illustrates a computerized system for forming and analyzing arrays of biological materials such as RNA or DNA. A computer 100 is used to design arrays of biological polymers such as RNA or DNA. The computer 100 may be, for example, an appropriately programmed IBM personal computer compatible running Windows NT including appropriate memory and a CPU as shown in Figs. 1 and 2. The computer system 100 obtains inputs from a user regarding characteristics of a gene of interest, and other inputs regarding the desired features of the array. Optionally, the computer system may obtain information regarding a specific genetic sequence of interest from an external or internal database 102 such as GenBank. The output of the computer system 100 is a set of chip design computer files 104 in the form of, for example, a switch matrix, as described in PCT application WO 92/10092, and other associated computer files.

The chip design files are provided to a system 106 that designs the lithographic masks used in the fabrication of arrays of molecules such as DNA. The system or process 106 may include the hardware necessary to manufacture masks 110 and also the necessary computer hardware and software 108 necessary to lay the mask patterns out on the mask in an efficient manner. As with the other features in Fig. 3, such equipment may or may not be located at the same physical site, but is shown together for ease of illustration in Fig. 3. The system 106 generates masks 110 or other synthesis patterns such as chrome-on-glass masks for use in the fabrication of polymer arrays.

The masks 110, as well as selected information relating to the design of the chips from system 100, are used in a synthesis system 112. Synthesis system 112 includes the necessary hardware and software used to fabricate arrays of polymers on a substrate or chip 114. For example, synthesizer 112 includes a light source 116 and a chemical flow cell 118 on which the substrate or chip 114 is placed. Mask 110 is placed

between the light source and the substrate/chip, and the two are translated relative to each other at appropriate times for deprotection of selected regions of the chip. Selected chemical reagents are directed through flow cell 118 for coupling to deprotected regions, as well as for washing and other operations. All operations are preferably directed by an appropriately programmed computer 119, which may or may not be the same computer as the computer(s) used in mask design and mask making.

The substrates fabricated by synthesis system 112 are optionally diced into smaller chips and exposed to marked targets. The targets may or may not be complementary to one or more of the molecules on the substrate. The targets are marked with a label such as a fluorescein label (indicated by an asterisk in Fig. 3) and placed in scanning system 120. Scanning system 120 again operates under the direction of an appropriately programmed digital computer 122, which also may or may not be the same computer as the computers used in synthesis, mask making, and mask design. The scanner 120 includes a detection device 124 such as a confocal microscope or CCD (charge-coupled device) that is used to detect the location where labeled target has bound to the substrate. The output of scanner 120 is an image file(s) 124 indicating, in the case of fluorescein labeled target, the fluorescence intensity (photon counts or other related measurements, such as voltage) as a function of position on the substrate. Since higher photon counts will be observed where the labeled target has bound more strongly to the array of polymers, and since the monomer sequence of the polymers on the substrate is known as a function of position, it becomes possible to determine the sequence(s) of polymer(s) on the substrate that are complementary to the target.

The image file 124 is provided as input to an analysis system 126 that incorporates the visualization and analysis methods of the present invention. Again, the analysis system may be any one of a wide variety of computer system. The present invention provides various methods of analyzing and visualizing the chip design files and the image files, providing appropriate output 128. The chip design need not include any particular number of probes. It should be understood that the present invention does not require any particular source of expression level information.

Fig. 4 provides a simplified illustration of the overall software system used in the operation of one embodiment of the invention. As shown in Fig. 4, the system first identifies the nucleotide sequence(s) or targets that would be of interest in a particular expression level analysis at step 202. The sequences of interest correspond to

mRNA transcripts of one or more genes, ESTs or nucleic acids derived from the mRNA transcripts. Sequence selection may be provided via manual input of text files or may be from external sources such as GenBank.

At step 204 the system evaluates the sequences of interest to determine or assist the user in determining which probes would be desirable on the chip, and provides an appropriate "layout" on the chip for the probes. The process of selecting probes for an expression level analysis is explained in PCT Publication No. WO 97/10365, the contents of which are herein incorporated by reference. An alternative probe selection process that does not require prior knowledge of sequences of interest is explained in PCT Publication No. WO97/27317 (Attorney Docket No. 18547-019410PC), the contents of which are herein incorporated by reference. Further general background on probe selection is found in PCT Publication No. WO95/11995 (Attorney Docket No. 18547-004111PC) and PCT Publication No. WO97/29212 (Attorney Docket No. 18547-018540PC), the contents of which are herein incorporated by reference. The term "perfect match probe" refers to a probe that has a sequence that is perfectly complementary to a particular target sequence. The test probe is typically perfectly complementary to a portion (subsequence) of the target sequence. The term "mismatch control" or "mismatch probe" refer to probes whose sequence is deliberately selected not to be perfectly complementary to a particular target sequence. For each mismatch (MM) control in an array there typically exists a corresponding perfect match (PM) probe that is perfectly complementary to the same particular target sequence.

The process compares hybridization intensities of pairs of perfect match and mismatch probes that are preferably covalently attached to the surface of a substrate or chip. Most preferably, the nucleic acid probes have a density greater than about 60 different nucleic acid probes per 1 cm^2 of the substrate.

Initially, nucleic acid probes are selected that are complementary to the target sequence. These probes are the perfect match probes. Another set of probes is specified that are intended to be not perfectly complementary to the target sequence. These probes are the mismatch probes and each mismatch probe includes at least one nucleotide mismatch from a perfect match probe. Accordingly, a mismatch probe and the perfect match probe to which it is identical except for one base make up a pair. As mentioned earlier, the nucleotide mismatch is preferably near the center of the mismatch probe.

The probe lengths of the perfect match probes are typically chosen to exhibit detectably greater hybridization with the target sequence relative to the mismatch probes. For example, the nucleic acid probes may be all 20-mers. However, probes of varying lengths may also be synthesized on the substrate for any number of reasons including resolving ambiguities.

Again referring to Fig. 4, at step 206 the masks for the synthesis are designed. At step 208 the software utilizes the mask design and layout information to make the DNA or other polymer chips. This step 208 will control, among other things, relative translation of a substrate and the mask, the flow of desired reagents through a flow cell, the synthesis temperature of the flow cell, and other parameters. At step 210, another piece of software is used in scanning a chip thus synthesized and exposed to a labeled target. The software controls the scanning of the chip, and stores the data thus obtained in a file that may later be utilized to extract hybridization information.

At step 212 a computer system utilizes the layout information and the fluorescence information to evaluate the hybridized nucleic acid probes on the chip. Among the important pieces of information obtained from DNA chips are the relative fluorescent intensities obtained from the perfect match probes and mismatch probes. These intensity levels are used to estimate an expression level for a gene or EST. The computer system used for analysis will preferably have available other details of the experiment including possibly the gene name, gene sequence, probe sequences, probe locations on the substrate, and the like.

According to the present invention, at step 214, the same computer system used for analysis or another one displays the expression level information in a format useful for identifying genes of interest. The visualized expression level information may include information collected from multiple applications of one or more previous steps of Fig. 4.

Fig. 5 is a flowchart describing steps of estimating an expression level for a particular gene and determining whether the expression level is sufficiently high to be displayed. At step 952, the computer system receives raw scan data of N pairs of perfect match and mismatch probes. In a preferred embodiment, the hybridization intensities are photon counts from a fluorescein labeled target that has hybridized to the probes on the substrate. For simplicity, the hybridization intensity of a perfect match probe will be designed "I_{pm}" and the hybridization intensity of a mismatch probe will be designed

"I_{mm}."

Hybridization intensities for a pair of probes are retrieved at step 954. The background signal intensity is subtracted from each of the hybridization intensities of the pair at step 956. Background subtraction can also be performed on all the raw scan data at the same time.

At step 958, the hybridization intensities of the pair of probes are compared to a difference threshold (D) and a ratio threshold (R). It is determined if the difference between the hybridization intensities of the pair (I_{pm} - I_{mm}) is greater than or equal to the difference threshold AND the quotient of the hybridization intensities of the pair (I_{pm} / I_{mm}) is greater than or equal to the ratio threshold. The difference thresholds are typically user defined values that have been determined to produce accurate expression monitoring of a gene or genes. In one embodiment, the difference threshold is 20 and the ratio threshold is 1.2.

If $I_{pm} - I_{mm} >= D$ and $I_{pm} / I_{mm} >= R$, the value NPOS is incremented at step 960. In general, NPOS is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely expressed. NPOS is utilized in a determination of the expression of the gene.

At step 962, it is determined if $I_{mm} - I_{pm} >= D$ and $I_{mm} / I_{pm} >= R$. If these expressions are true, the value NNEG is incremented at step 964. In general, NNEG is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely not expressed. NNEG, like NPOS, is utilized in a determination of the expression of the gene.

For each pair that exhibits hybridization intensities either indicating the gene is expressed or not expressed, a log ratio value (LR) and intensity difference value (IDIF) are calculated at step 966. LR is calculated by the log of the quotient of the hybridization intensities of the pair (I_{pm} / I_{mm}). The IDIF is calculated by the difference between the hybridization intensities of the pair (I_{pm} - I_{mm}). If there is a next pair of hybridization intensities at step 968, they are retrieved at step 954.

At step 972, a decision matrix is utilized to indicate if the gene is expressed. The decision matrix utilizes the values N, NPOS, NNEG, LR (multiple LRs), and IDIF (multiple IDIFs). The following four assignments are performed:

$$P1 = NPOS / NNEG$$

$$P2 = NPOS / N$$

P3 = SUM(LR) / N

P4 = SUM(IDIF)/N

These P values are then utilized to determine if the gene is expressed and if the expression level should be displayed. In a preferred embodiment, the expression level of a gene should be displayed if:

P1 > 2.2

P2 > 0.3

P3 > 0.8

P4 > 30

Once all the pairs of probes have been processed and the expression of the gene indicated, an average of the IDIF values for the probes that incremented NPOS or NNEG is calculated at step 975, which is utilized as an expression level. Of course, other values including one of P1 through P4 could be used to indicate expression level.

For simplicity, Fig. 5 was described in reference to a single gene or EST. However, the visualization system of the present invention displays expression results for many genes to facilitate discovery of genes of interest or ESTs. Furthermore, the present invention contemplates display of expression levels of a single gene or ESTs as collected from two or more different samples such as tissue samples. The sample sources preferably differ in some characteristic. It will be understood that when the term "sample" is used herein, measurements made on a single "sample" can be based on an aggregation of multiple sample collection events or even multiple organisms.

Fig. 6 shows a screen display illustrating gene expression levels for multiple genes as collected from two tissue samples. A displayed horizontal axis 1002 represents expression level measured in one or more nucleic acid samples taken from the first tissue sample. A displayed vertical axis 1004 represents expression level in one or more nucleic acid samples taken from the second tissue sample. Each of marks 1006 represent a particular gene whose expression level has been measured in both the first and second tissue samples. Each mark 1006 is placed at a distance from vertical axis 1004 corresponding to expression level in the first tissue sample and at a distance from the horizontal axis 1002 corresponding to expression level in the second tissue sample.

The expression levels used for determining the position of marks 1006 are preferably taken from the result of step 975. The position of each of marks 1006 depends on two iterations of the steps of Fig. 5, once for the sample taken from the first tissue

sample and once for the sample taken from the second tissue sample. However, a mark is preferably displayed only if one of the samples meets the threshold criteria at step 972.

In the depicted representative screen display, the first tissue sample is a cancerous tissue sample and the second tissue sample is a normal tissue sample. The individual marks represent the expression levels of selected genes in both cancerous and normal tissue. A first group of marks 1008 represent genes that are neither tumor suppressors nor oncogenes since their expression levels are roughly similar for both normal and cancerous tissue. These marks 1008 fall roughly along a line which is rotated 45 degrees from each of the axes. A second group of marks 1010 represent genes that are likely oncogenes since their expression levels are found to be significantly higher in cancerous tissue than in normal tissue. A third group of marks 1012 represent genes that are likely tumor suppressors since their expression levels are found to be significantly higher in normal tissue than in cancerous tissue. It will be appreciated that expression levels for large numbers of genes can be reviewed at once to discover the oncogenes and tumor suppressors.

Although in the depicted display, the two types of tissue are normal tissue and cancerous tissue, the present invention would aid in the discovery of genes whose expression is associated with any characteristic that varies among tissue samples. For example, once can compare expression results from tissue from individuals who have been exposed to HIV but remain infected to tissue obtained from infected individuals to identify genes conferring resistance to HIV. One can compare expression results between tissue from plants that survive drought to plants that do not. One can compare expression levels among tissue samples at successive stages or severity levels of the same disease, among tissue samples where different ultimate outcomes of the disease (e.g., patient death or remission) are known, among diseased tissue samples that have been subject to different treatment regimes including e.g, chemotherapy, antisense RNA, etc. For cancers, one can compare expression levels between malignant cells and non-malignant cells. Also expression levels can be compared among different organs, between species, and among different stages of development of an organ.

It will be appreciated that the present invention also encompasses displays with more than two dimensions. A third visual dimension can be used to illustrate expression level from a third tissue sample. The time dimension can also be used to illustrate successive groups of two or three tissue samples at successive time periods.

The time dimension can be also used to correspond to tissue samples obtained at, e.g, successive stages of a disease.

Other interface methods corresponding to human senses other than sight can also be incorporated within the presentation system of the present invention. The senses may correspond to additional dimensions. For example, marks can be displayed in succession accompanies by a sound having characteristics corresponding to expression level in another tissue sample.

The user can employ a cursor 1014 to identify a particular mark as being of interest. Cursor 1014 can be moved to a particular mark by use of, e.g., mouse 11. Once cursor 1014 is over a mark of interest, the mark can be selected by, e.g., depression of one of mouse buttons 13. Selection of a particular mark can be facilitated by use of a zoom display feature (not shown). Once a particular mark is selected, further information is displayed about the gene represented by the mark. A special mouse can transmit a tactile sensation back to the user corresponding to expression level in a tissue sample as the user passes the mouse over a corresponding mark.

It will be appreciated that the display of Fig. 6 is not limited to expression information. The two dimensions of Fig. 6 may correspond to indicators of the presence of various polymers other than nucleic acids in two different samples. For example, each mark may correspond to a different polymer, polypeptide, or other compound. The distance of the mark from each axis would correspond to a measure of presence of the particular polymer in the sample corresponding to the axis. One possible measure is produced by fluorescently tagging polymer samples such as protein samples and exposing a probe array such as a peptide probe array to the protein samples. The fluorescent intensity of the probes will then correspond to the bonding affinity of the sample to the probes. The intensity measurement or a measurement derived from the intensity measurement may then be used to position the marks of Fig. 6.

Fig. 7A shows a screen display giving information about a particular gene selected from the display of Fig. 6. A cluster number 702, a GenBank accession number 704, and a verbal description 706 for the selected gene are displayed. The user can also select a number of marks 1006 by circling them with cursor 1014. Then a list of information as shown in Fig. 7A is displayed for all the genes corresponding to the selected marks.

By selecting GenBank accession number 704 with another cursor (not

shown), the user can direct retrieval of the GenBank information for the selected gene. If the GenBank information is not available locally, the retrieval process can include formulating a query and transmitting the query to a GenBank web site. Once the GenBank information is retrieved, it can also be displayed. Fig. 7B depicts the GenBank information for the gene identified in Fig. 7A.

In the foregoing specification, the invention has been described with reference to specific exemplary embodiments thereof. It will, however, be evident that various modifications and changes may be made thereunto without departing from the broader spirit and scope of the invention as set forth in the appended claims and their full scope of equivalents.